Computational Biology and Bioinformatics
2015; 3(6): 88-94
Published online December 30, 2015 (http://www.sciencepublishinggroup.com/j/cbb)
doi: 10.11648/j.cbb.20150306.12
ISSN: 2330-8265 (Print); ISSN: 2330-8281 (Online)

Methods for Evaluating Agglomerative Hierarchical Clustering for Gene Expression Data: A Comparative Study

Md. Bipul Hossen[1, *], Md. Siraj-Ud-Doulah[1], Aminul Hoque[2]

[1]Department of Statistics, Begum Rokeya University, Rangpur, Bangladesh
[2]Department of Statistics, Rajshahi University, Rajshahi, Bangladesh

Email address:
mbipu.ru@gmail.com (Md. B. Hossen), sdoulah_brur@yahoo.com (Md. Siraj-Ud-Doulah), mdaminulh@gmail.com (A. Hoque)

To cite this article:
Md. Bipul Hossen, Md. Siraj-Ud-Doulah, Aminul Hoque. Methods for Evaluating Agglomerative Hierarchical Clustering for Gene Expression Data: A Comparative Study. *Computational Biology and Bioinformatics.* Vol. 3, No. 6, 2015, pp. 88-94. doi: 10.11648/j.cbb.20150306.12

Abstract: Microarray is already well established techniques to understand various cellular functions by profiling transcriptomics data. To capture the overall feature of high dimensional variable datasets in microarray data, various analytical and statistical approaches are already developed. One of the most widely used Agglomerative Hierarchical Clustering (AHC) methods is the cluster analysis of gene expression data; however, little work has been done to compare the performance of clustering methods on gene expression data, where some authors used three or four AHC methods and some others used at most five AHC methods. All of the authors concretely suggested complete linkage method to further researchers to determine the best method for clustering their gene expression data. This paper compared the performance of seven AHC methods for clustering gene expression data with respect to five major proximity measures. We used corrected Rand (cR) Index to compare the performance of each clustering method. To illustrate the results, we found that the clustering method Ward exhibited the best performance among all of the AHC methods as well as the proximity measure Cosine performed better in comparison to all the other measures in both type of Affymetrix and cDNA datasets.

Keywords: Agglomerative Hierarchical Clustering, Proximity Measures, Corrected Rand Index, Gene Expressions Data

1. Introduction

Cluster analysis programs are routinely run as a first step of data summary and grouping genes in a microarray data analysis. There are many clustering methods, such as hierarchical clustering method, which can classify into agglomerative hierarchical methods and divisive hierarchical methods [28, 18]. Agglomerative Hierarchical Clustering (AHC) process starts with these single observation clusters and progressively combines pairs of clusters, forming smaller numbers of clusters that contain more observations [17, 29]. Several AHC methods are well established [5, 11]. It is essential to know which clustering method is best for which type of microarray gene (cancer) data. Microarray gene expression data allow us to quantitatively and simultaneously monitor the expression of thousands of genes under different conditions [1, 3]. DNA microarray technology has now made it possible to simultaneously monitor the expression levels of thousands of genes during important biological processes and across collections of related samples.

Generally the gene expression microarray technology is available in two types of platforms, single channel microarrays (Affymetrix) and double channel microarrays (cDNA) [2, 4]. One of the characteristics of gene expression data is that it is meaningful to cluster both genes and samples [13, 27]. Therefore there are two types of gene expression data clustering: gene based clustering and sample based clustering. In sample based clustering, samples are treated as objects while genes are treated as features and samples are partitioned into homogeneous groups [12, 19]. The goal of sample-based clustering is to identify the phenotype structures or substructures of the samples. This study conducted only sample based clustering.

There are a small number of analyses in literature for evaluating the performance of different clustering method applied to gene expression data. Three AHC methods (Single Linkage, Complete Linkage and Average Linkage) were used to identify the clustering performance in gene expression data [7, 8, 16, 25]. Four AHC methods (Single Linkage, Complete Linkage, Average Linkage and Centroid Linkage) were practiced to evaluate the clustering performance in their

Table 1. Several Proximity Measures.

Methods	Descriptions	Functions
Euclidean	It is the square root of the sum of squared differences between corresponding elements of the two vectors.	$d(x,y) = \sqrt{\sum_{i}^{m}(x_i - y_i)^2}$
Manhattan	Measures distance following only axis-aligned directions	$d(x,y) = \sum_{i=1}^{m} \lvert x_i - y_i \rvert$
Pearsons Correlation	Measures the similarity between the shapes of two expression patterns (profiles)	$d(x,y) = 1 - \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2 \sum_{i=1}^{m}(y_i - \bar{y})^2}}$
Spearman Correlation	Measures the degree of a monotonic relationship between two variables, without making any assumptions about the frequency distribution of the variables.	$d(x,y) = 1 - \frac{\sum_{i=1}^{m}(\acute{x}_\imath - \bar{\acute{x}})(\acute{y}_\imath - \bar{\acute{y}})}{\sqrt{\sum_{i=1}^{m}(\acute{x}_\imath - \bar{\acute{x}})^2 \sum_{i=1}^{m}(\acute{y}_\imath - \bar{\acute{y}})^2}}$
Cosine Correlation	Measure of similarity of two non-binary vectors.	$d(x,y) = 1 - \frac{\acute{x}y}{\lVert x \rVert \lVert y \rVert} = 1 - \frac{\lvert \sum_{i=1}^{m} x_i y_i \rvert}{\sqrt{\sum_{i=1}^{m} {x_i}^2 \sum_{i=1}^{m} {y_i}^2}}$

analysis [6, 9, 10]. Five AHC methods were also compared to check better clustering methods in their datasets [14, 15]. However all of the author's demonstrated complete linkage isbetter measure to evaluate gene expression data in their analysis.

1.1. Distance and Similarity Measures for Gene Expression Data

Distances and similarities play an important role in cluster analysis [23, 26]. In this section, we introduce some distance and similarity measures for gene expression data in Table 1. In shortly discuss the distance and similarity measures for gene expression data, we start with some notation. Let $\mathrm{x} = (\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{m})^\mathrm{T}$ and $\mathrm{y} = (\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_\mathrm{m})^\mathrm{T}$ be two numerical vectors that denote two gene expression data objects, where the objects can be either genes or samples and m is the number of features [20, 21, 22].

1.2. Checking Validity of Clusters

Clustering is an unsupervised process in the data mining and pattern recognition and most of the clustering methods are very sensitive to their input parameters. Therefore it is very important to evaluate the result of the clustering methods. It is difficult to define when a clustering result is acceptable, thus several clustering validity techniques have been developed. In this study the most commonly used validity techniques as Corrected Rand Index are used.

1.3. Corrected Rand (cR) Index

Measuring the efficiency of the AHC methods in recovering the true partition of the data sets we use the corrected Rand index [23, 24]. The corrected Rand index takes values from -1 to 1, with 1 indicating a perfect agreement between the partitions and values near 0 or negatives corresponding to cluster agreement found by chance. Unlike the majority of other indices, the corrected Rand is not biased towards a given method or number of cluster in the partition [4, 23]. Given a set S of n elements and two groupings (e.g. clustering's) of these points, namely x={X_1,X_2...,X_R} and y={Y_1,Y_2,...,Y_S}, the overlap between X and Y can be summarized in a contingency table $[n_{ij}]$ where each entry n_{ij} denotes the number of objects in common X_I and Y_J: $n_{ij} = |X_i \cap Y_j|$ The corrected form of Rand Index is cR and the index is given as

$$cR = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}$$

where $n_{ij,}a_i, and\ b_j$are values from the contingency table.

Motivated by this problem it is important to consider all of the methods for gene expression data by assessing which method are comparatively best. This paper tries to compare seven AHC methods which are single linkage (cluster separation as distance between two nearest objects), complete linkage (as previously, but two furthest objects), average linkage (average distance between all pairs), centroid (distance between centroid's of each cluster), Ward's method (minimizes ANOVA Sum of Squared Errors between two clusters) median (the similarity is based on the distance between the two medians) and mcquitty (Average the distances from both parts of the new cluster) in both Affymetrix and cDNA datasets. It is also provided a detailed graphical and analytical comparison of seven agglomerative hierarchical clustering (AHC) methods and five proximity measures. We used Bar diagram as well as Box and Whisker plot with respect to Corrected Rand Index to check the suitable AHC method for clustering. In this paper the AHC algorithm with different linkages and several proximity measures are implemented using language programming R 3.0.2 with mclust and proxy packages. Several times Ms-Excel and Ms-Word are used as calculation and typing software.

2. Experiments and Results

Thirty three publicly available microarray data sets are included in our analysis [25]. These data sets were obtained using two microarrays technologies: single channel Affymetrix chips (21 sets) and double-channel cDNA (12 sets). We compare seven different types of clustering methods with regard five proximity measures. Mainly the gene expression data is so much noisy, mixture with expression pattern, down regulated and up regulated so it is necessary to take preprocess before differential expression analysis. To adjust data for technical variation, as opposed to biological differences between the samples we have preprocessed only Affymetrix data by using standardization technique. It is mentioned that the cDNA datasets were preprocessed. The experimental datasets are given in Table 2.

At first we present some graphical displays for both gene expression datasets. For each of the seven AHC methods of clustering, we represent the results by using Bar diagram, Box and whisker plot to compare which AHC methods is best and the graph are given in Figure 1, Figure 2 and Figure 3.

The mean values of the corrected Rand (cR) index of the experiments with Affymetrix 21 datasets are presented in Figure 1. The ward method obtained the highest value with respect to proximity measures when compared to those achieved by the other methods, whereas the second best method, complete linkage, which is one of most traditionally used method obtained the lowest values in comparison to all the other methods.

Figure 2 illustrates the mean values of the corrected Rand for the experiments performed with the cDNA 12 datasets. The ward method achieved the highest value with respect to proximity measures in comparison to all the other methods. The median and the centroid methods attained the lowest values in comparison to all the other methods.

Table 2. *Data Description.*

Dataset	Chip	Tissue	n	#c	Dist. Classes	m	d
Armstrong-V1 [52]	Affy	Blood	72	2	24,48	12582	1081
Armstrong-V2 [52]	Affy	Blood	72	3	24,20,28	12582	2194
Bhattacharjee [9]	Affy	Lung	203	5	139,17,6,21,20	12600	1543
Chowdary [13]	Affy	Breast, Colon	104	2	62,42	22283	182
Dyrskjot [14]	Affy	Bladder	40	3	9,20,11	7129	1203
Golub-V1 [3]	Affy	Bone marrow	72	2	47,25	7129	1877
Golub-V2 [3]	Affy	Bone marrow	72	3	38,9,25	7129	1877
Gordon [53]	Affy	Lung	181	2	31,150	12533	1626
Laiho [15]	Affy	Colon	37	2	8,29	22883	2202
Nutt-V1 [54]	Affy	Brain	50	4	14,7,14,15	12625	1377
Nutt-V2 [54]	Affy	Brain	28	2	14,14	12625	1070
Nutt-V3 [54]	Affy	Brain	22	2	7,15	12625	1152
Pomeroy-V1 [55]	Affy	Brain	34	2	25,9	7129	857
Pomeroy-V2 [55]	Affy	Brain	42	5	10,10,10,4,8	7129	1379
Ramaswamy [50]	Affy	Multi-tissue	190	14	11,10,11,11,22,10,11,10,30,11,11,11,11,20	16063	1363
Shipp [56]	Affy	Blood	77	2	58,19	7129	798
Singh [19]	Affy	Prostate	102	2	58,19	12600	339
Su [57]	Affy	Multi-tissue	174	10	26,8,26,23,12,11,7,27,6,28	12533	1571
West [58]	Affy	Breast	49	2	25,24	7129	1198
Yeoh-V1 [20]	Affy	Bone marrow	248	2	43,205	12625	2526
Yeoh-V2 [20]	Affy	Bone marrow	248	6	15,27,64,20,79,43	12625	2526
Alizadeh-V1 [4]	cDNA	Blood	42	2	21,21	4022	1095
Alizadeh-V2 [4]	cDNA	Blood	62	3	42,9,11	4022	2093
Alizadeh-V3 [4]	cDNA	Blood	62	4	21,21,9,11	4022	2093
Bittner [10]	cDNA	Skin	38	2	19,19	8067	2201
Bredel [11]	cDNA	Brain	50	3	31,14,5	41472	1739
Chen [12]	cDNA	Liver	180	2	104,76	22699	85
Garber [59]	cDNA	Lung	66	4	17,40,4,5	24192	4553
Khan [60]	cDNA	Multi-tissue	83	4	29,11,18,25	6567	1069
Liang [17]	cDNA	Brain	37	3	28,6,3	24192	1411
Risinger [18]	cDNA	Endometrium	42	4	13,3,19,7	8872	1771
Tomlins-V1 [61]	cDNA	Prostate	104	5	27,20,32,13,12	20000	2315
Tomlins-V2 [61]	cDNA	Prostate	92	4	27,20,32,13	20000	1288

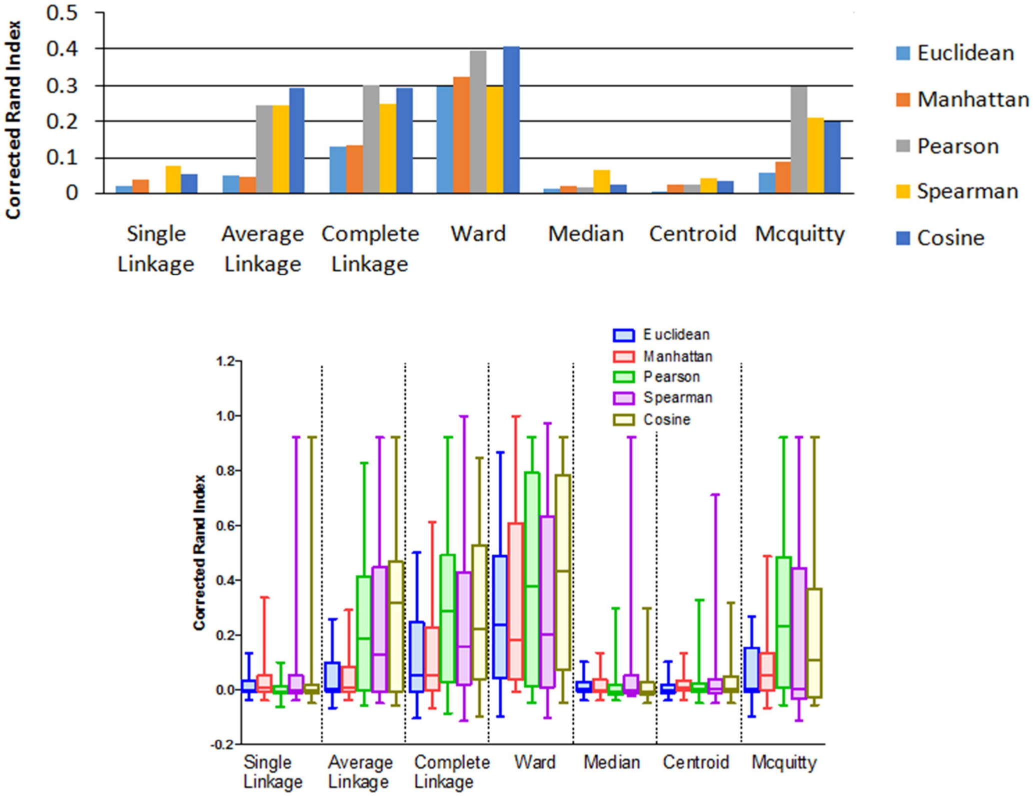

Figure 1. *Several Agglomerative Hierarchical Clustering Methods with respect to Proximity measures of Affymetrix datasets.*

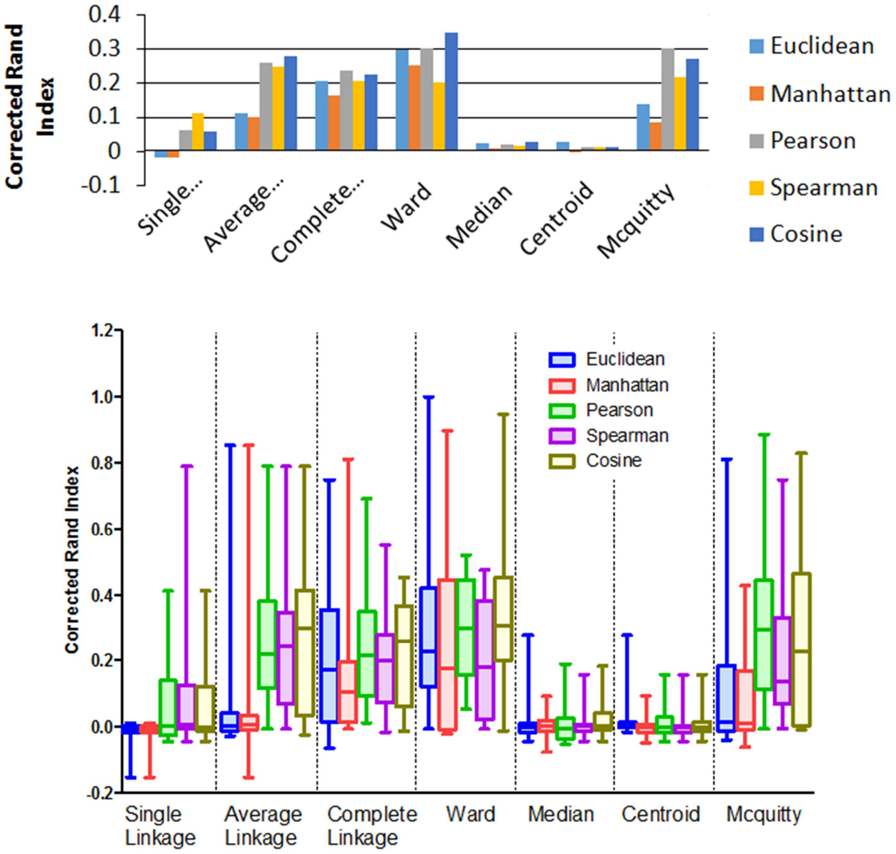

Figure 2. *Several Agglomerative Hierarchical Clustering Methods with respect to Proximity measures of cDNA datasets*

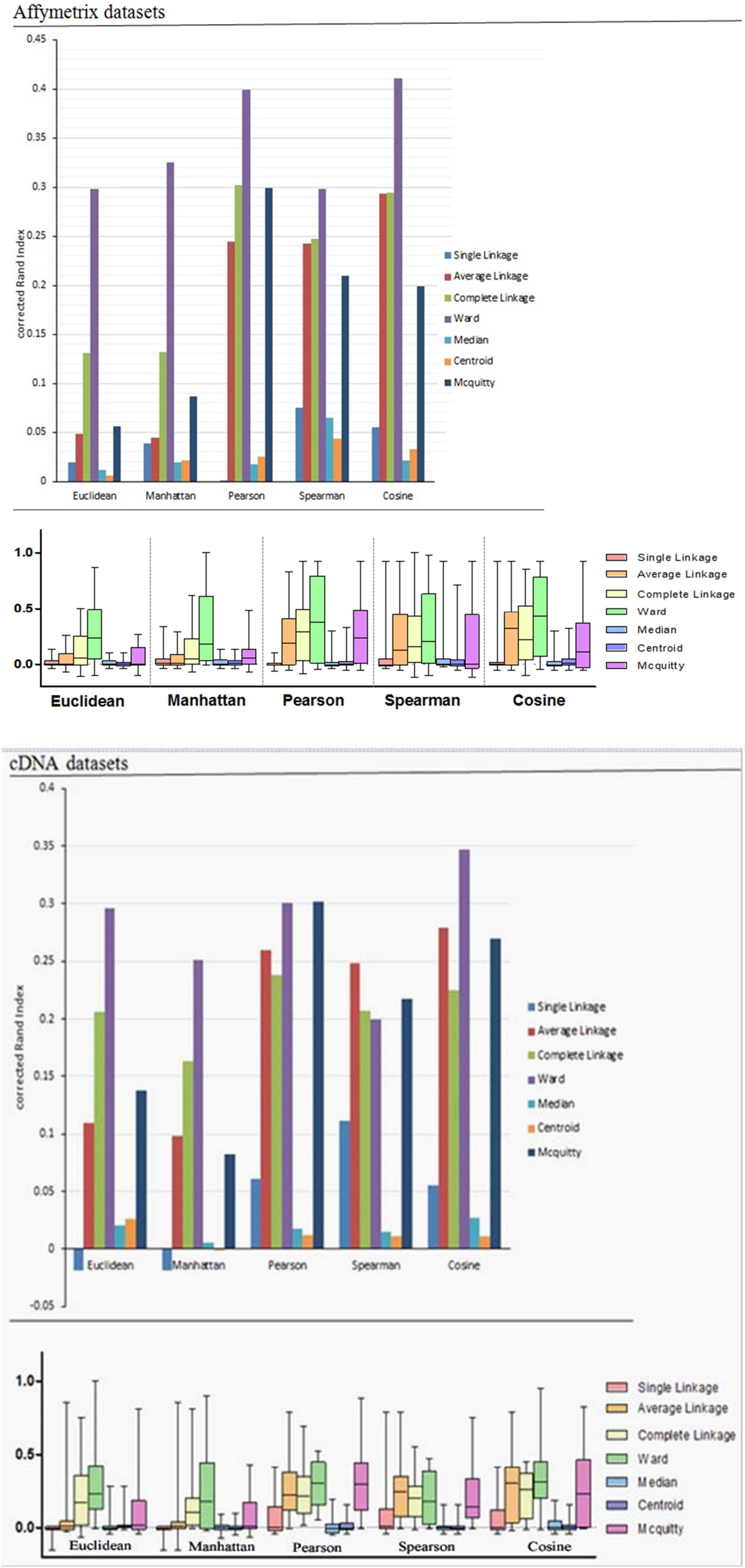

Figure 3. *Performance of several proximity measures of both datasets.*

In another kind of analysis, we also investigated the performance of proximity measures corresponding to the AHC methods. The mean values of the corrected Rand for the experiments performed with the Affymetrix and cDNA datasets are presented in Figure 3. Based on this Figure 3, we found that the Cosine measures achieved the highest value in compared to the other measures.

***Table 3.** The mean corrected Rand value of Affymetrix and cDNA datasets.*

Affymetrix datasets					
	Euclidean	**Manhattan**	**Pearson**	**Spearman**	**Cosine**
SL	0.0195	0.0383	0.0013	0.0753	0.0546
AL	0.0480	0.0441	0.2442	0.2427	0.2934
CL	0.1307	0.1320	0.3022	0.2470	0.2945
Ward	**0.2983**	**0.3251**	**0.3987**	**0.2981**	**0.4103**
Median	0.0119	0.0196	0.0171	0.0641	0.0213
Centroid	0.0055	0.0215	0.0250	0.04304	0.0330
Mcquitty	0.0555	0.0864	0.2988	0.2095	0.1990
cDNA datasets					
SL	-0.0190	-0.0188	0.0612	0.1114	0.0549
AL	0.1093	0.0981	0.2598	0.2481	0.2790
CL	0.2061	0.1632	0.2377	0.2070	0.2252
Ward	**0.2960**	**0.2511**	**0.3004**	**0.1997**	**0.3465**
Median	0.0204	0.0059	0.0178	0.0153	0.0271
Centroid	0.0261	-0.0008	0.0119	0.0113	0.0108
Mcquitty	0.1377	0.0824	0.3019	0.2173	0.2699

[N. B.: SL= Single Linkage, CL= Complete Linkage, AL= Average Linkage]

In terms of results for both Affymetrix and cDNA datasets, Table 3showed the average corrected Rand index values of seven Agglomerative Hierarchical Clustering methods with respect to proximity measures. We observed that ward method performed better than any other methods due to achieving the highest cR values. Furthermore, the cosine proximity measures showed the highest cR values in comparison to all the other proximity measures.

3. Conclusion

Cluster analysis techniques of gene expression microarray data is of increasing interest in the field of functional genomics. One of the reasons for this is the need for molecular-based refinement of broadly defined biological classes, with implications in cancer diagnosis, prognosis and treatment. For this reason, we revisited two types of microarray datasets: Affymetrix and cDNA. This paper shows a comparative study of seven AHC methods regarding to five proximity measures applied in a large scale datasets. The corrected Rand (cR) index was used to calculate the accuracy of the clustering. We found that the performance of Ward method is superior to all other methods for both types of datasets. We also found that the performance of Cosine is better than all other proximity measures for two types of datasets. It is recommended that Ward method with cosine distance are used to analyze Affymetrix and cDNA gene expression datasets.

Acknowledgement

The authors would like to thank the anonymous reviewers for their helpful comments.

References

[1] Brown M P and Bostein D (1999); Exploring the new world of genome with DNA microarrays. *Nature Genetics*, 21: 33-37.

[2] Quackenbush J (2001); Computational analysis of cDNA microarray data. *Nature Reviews*. 6(2):418-428.

[3] Slonim D (2002); From patterns to pathways: gene expression data analysis comes of age. *Nature Genetics*. 32:502-508.

[4] Monti S, Tamayo P, Mesirov J, Golub T (2003); Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data; *Machine Learning*. 52:91-118.

[5] Cunningham K M and Ogilvie J C (1972); Evaluation of hierarchical grouping techniques: A preliminary study. *The Computer Journal*, 15: 209–213.

[6] Hubert L (1974); Approximate evaluation techniques for the single-link and complete link hierarchical clustering procedures. *Journal of the American Statistical Association*, 69, 698–704.

[7] Baker F B (1974); Stability of two hierarchical grouping techniques – Case I: Sensitivity to data errors. *Journal of the American Statistical Association*, 69: 440–445.

[8] Kuiper F K and Fisher L (1975); A Monte Carlo comparison of six clustering procedures. *Biometrics*, 31: 777–783.

[9] Blashfield R K (1976); Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. *The Psychological Bulletin*, 83: 377–388.

[10] Hands S and Everitt B (1987); A Monte Carlo study of the recovery of cluster structure in binary data by hierarchical clustering techniques. *Multivariate Behavioral Research*, 22: 235–243.

[11] Johnson R A and Wichern D W (2002); *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ: Prentice Hall.

[12] Jaskowiak P A, Campello R J G B and Costa I G (2013); Proximity Measures for Clustering Gene Expression Microarray Data: A Validation Methodology and a Comparative Analysis, *Computational Biology and Bioinformatics*. 10 (4):845-857.

[13] Costa I G, Carvalho F A D and Souto M C P D (2004); Comparative Analysis of Clustering Methods for Gene Expression Time Course Data. *Genetics and Molecular Biology*, 27: 4623-4631.

[14] Datta S and Datta S (2006); Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes.*BMC Bioinformatics*, 7: 397.

[15] Kerr G, Ruskin H J, Crane M and Doolan P (2008); Techniques for clustering gene expression data. *ComputBiol Med*, 38(3): 283-293.

[16] Geetha T and Michael A (2010); Enhanced Hierarchical Clustering for Gene Expression data. *International Journal of Computer Applications* 1(20):92–98.

[17] Milligan G W (1980); An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, 45: 325–342.

[18] Sasirekha K and Baby P (2013); Agglomerative Hierarchical Clustering Algorithm-A Review. *International Journal of Scientific and Research Publications*, 3(3):01-03.

[19] Frakes W B and Baeza-Yates R (1992); *Information Retrieval: Data Structures and Algorithms*, Upper Saddle River, NJ: Prentice Hall.

[20] Guojun G, Chaoqun M and Jianhong W (2007); *Data Clustering: Theory, Algorithms, and Applications*, ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA.

[21] Gentleman R, Ding B, Dudoit S and Ibrahim J (2005); Bioinformatics and Computational Biology Solutions Using R and Bioconductor Statistics for Biology and Health, 189-208.

[22] Pablo A Jaskowiak, Ricardo J G B Campello and Ivan G Costa (2013); Proximity Measures for Clustering Gene Expression Microarray Data: A Validation Methodology and a Comparative Analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(4):845-857.

[23] Jain A K and Dubes R C (1988); Algorithms for clustering data Prentice Hall.

[24] Milligan G W and Cooper M C (1988); A study of standardization of variables in cluster analysis. *Journal of Classification*, 5:181-204.

[25] Marcilio C P de Souto, Ivan G Costa, Daniel S A de Araujo, Teresa B Ludermir and Alexander Schliep (2008); Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics*, 01-14.

[26] Anderberg M (1973); Cluster analysis for applications. New York: Academic Press.

[27] Daxin J, Chun T, and Aidong Z (2004); Cluster Analysis for Gene Expression Data: A Survey, *IEEE Transactions on Knowledge and Data Engineering,* 16 (11):1370-1386.

[28] Eldesoky, A.E, M.Saleh, N.A. Sakr (2009); Novel Similarity Measure fo Document Clustering Basedon Topic Phrase, Interna-tional Conferenceon Networking and Media Convergence24: 92-96.

[29] Myatt, Glenn J,"Making Sense of Data II", 2009, Wiley, Canada.